Journal of Multivariate Analysis **79**, 1–32 (2001)
doi:10.1006/jmva.2000.1961, available online at http://www.idealibrary.com on IDEAL®

Semiparametric Mixtures in Case-Control Studies

S. A. Murphy[1]

Pennsylvania State University

and

A. W. van der Vaart[2]

Free University, Amsterdam, The Netherlands

Received May 7, 1997; published online June 11, 2001

control studies with a partially observed covariate. The likelihood is a combination of a nonparametric mixture, a parametric likelihood, and an empirical likelihood. We prove the asymptotic normality of the maximum likelihood estimator for the regression slope, the asymptotic chi-squared distribution of the likelihood ratio statistic, and the consistency of the observed information, in both the prospective and the retrospective model. © 2001 Academic Press

AMS 1991 subject classifications: 62G20; 62F12.

Key words and phrases: mixture model; maximum likelihood; asymptotic efficiency; semiparametric model; direct and indirect observations.

1. INTRODUCTION

In this paper we consider likelihood based inference in a class of models with partially observed covariates, with as a main example a logistic regression model for case-control studies considered by Roeder, Carroll, and Lindsay (1996). We prove the asymptotic normality of the semiparametric maximum likelihood estimator, obtain the asymptotic chi-squared distributions of the likelihood ratio statistics, and prove asymptotic consistency of the observed information.

The main example of the model is expressed in terms of a basic random vector (E, W, Z) whose distribution is described in the following way (our parameterization is slightly different from Roeder *et al.*, 1996):

[1] Research partially supported by NIDA Grant A P50 DA 10075-01.
[2] Research carried out while on leave at Université de Paris-sud.

0047-259X/01 $35.00

— E is a logistic regression on $\exp Z$ with intercept and slope β_0 and β_1, respectively;

— W is a a linear regression on Z with intercept and slope α_0 and α_1, respectively, and an $N(0, \sigma^2)$-error;

— Given Z the variables E and W are independent;

— Z has a completely unspecified distribution F on $\mathbb{R}$.

The approach of this paper applies to more general models. The regression structure (in our example linear on $\exp Z$ and Z, respectively) may be changed, and the prospective model introduced ahead does not require the presence of a 0–1 response, but could allow a general variable X instead of (E, W). Such variations might influence the conditions and technical arguments to carry through the proofs. Since a "super theorem" that would cover most cases of interest would be very complicated, we stick to the preceding set-up, but indicate in a final section which part of the results allow (easy) generalization. The parameter space for F is the set of non-degenerate probability distributions supported within a (known) compact interval $\mathscr{Z}$, which is in the real line for the preceding concrete example, but could be Euclidean in general. Since $\mathscr{Z}$ is assumed known, estimators of F will be constrained to have support within $\mathscr{Z}$. In our example, the parameter set for $\theta = (\alpha_0, \alpha_1, \beta_0, \beta_1, \sigma)$ is the set $\Theta = \mathbb{R}^4 \times (0, \infty)$.

Roeder *et al.* (1996) consider both a prospective and a retrospective (or case-control) model. In the prospective model we observe two independent random samples of sizes n_C and n_R from the distributions of (E, W, Z) and (E, W), respectively. (The indexes C and R are for "complete" and "reduced," respectively.) In the retrospective model we observe four independent random samples of sizes n_{C_0}, n_{C_1}, n_{R_0}, and n_{R_1} from the conditional distributions of (E, W, Z) given $E = 0$ and $E = 1$ and the conditional distributions of (E, W) given $E = 0$ and $E = 1$, respectively. (The extra indexes 1 and 0 are for "cases" and "controls," respectively.) In the terminology of Roeder *et al.* (1996), the covariate Z in a full observation (E, W, Z) is a "golden standard," but, in view of the costs of measurement, for a selection of observations only the "surrogate covariate" W is available. For instance, Z corresponds to the LDL-cholesterol and W to total cholesterol, and we are interested in heart disease $E = 1$.

The methods of the present paper apply to the case that the number of complete and reduced observations are of comparable magnitude. More precisely, we carry out asymptotics under the assumption that the fraction n_C/n_R is bounded away from 0 and ∞. If $n_R = 0$, then the model is purely parametric, and the classical results apply. If $n_C = 0$, then the model is a pure mixture model. To our knowledge, the behavior of likelihood based

procedures in the general mixture model is still unknown. (See Van der Vaart, 1996c, and Murphy and Van der Vaart, 1997, for some partial results.)

The model is semiparametric with a Euclidean parameter $\theta = (\alpha_0, \alpha_1, \beta_0, \beta_1, \sigma)$ and the unknown distribution F of the regression variable as the infinite-dimensional parameter. A density for the vector (E, W, Z) takes the form $p_\theta(e, w \mid z)\, dF(z)$ for, with ϕ denoting the standard normal density,

$$p_\theta(e, w \mid z) = \left(\frac{1}{1+\exp(-\beta_0-\beta_1 e^z)}\right)^e \left(\frac{\exp(-\beta_0-\beta_1 e^z)}{1+\exp(-\beta_0-\beta_1 e^z)}\right)^{1-e} \times \frac{1}{\sigma}\phi\left(\frac{w-\alpha_0-\alpha_1 z}{\sigma}\right).$$

To construct a likelihood function for statistical inference, we use the "empirical likelihood" for the distribution F of the observed Z_i, that is, we insert pointmasses $F\{Z_i\}$ in the likelihood. For the other part of the observations we use the density as a likelihood, as usual. This leads to the likelihoods for the prospective and retrospective models defined by, with $F\{z\}$ the measure of the point $\{z\}$,

$$\mathrm{Pros}(\theta, F) = \prod_{i=1}^{n_C} p_\theta(E_i, W_i \mid Z_i)\, F\{Z_i\} \prod_{i=n_C+1}^{n_C+n_R} \int p_\theta(E_i, W_i \mid s)\, dF(s),$$

$$\mathrm{Retro}(\theta, F) = \prod_{i=1}^{n_{C_0}} \frac{p_\theta(0, W_i \mid Z_i)\, F\{Z_i\}}{P_{\theta,F}(E=0)} \prod_{i=n_{C_0}+1}^{n_C} \frac{p_\theta(1, W_i \mid Z_i)\, F\{Z_i\}}{P_{\theta,F}(E=1)} \times \prod_{i=n_C+1}^{n_C+n_{R_0}} \frac{\int p_\theta(0, W_i \mid s)\, dF(s)}{P_{\theta,F}(E=0)} \prod_{i=n_C+n_{R_0}+1}^{n_C+n_R} \frac{\int p_\theta(1, W_i \mid s)\, dF(s)}{P_{\theta,F}(E=1)}.$$

Here $P_{\theta,F}(E=1) = \iint p_\theta(1, w \mid z)\, dF(z)\, dw$ is the probability that a randomly chosen subject from the population is a case.

In the prospective model the full parameter (θ, F) is identifiable. This model is closely related to a model introduced by Ibragimov and Hasminskii (1983), for which the maximum likelihood estimators where studied by Van der Vaart (1994, 1996a). Adapting and extending the methods developed in these papers, we shall show that the maximum likelihood estimator for θ is asymptotically normal. Furthermore, we study the likelihood ratio statistics for testing hypotheses concerning θ along the lines of Murphy and Van der Vaart (1997).

Roeder *et al.* (1996) have shown that in the retrospective model the parameter of prime interest, the logistic intercept β_1, is identifiable, and so are of course (α_0, α_1) and σ^2, but β_0 and F are confounded. They prove the following nice result.

LEMMA 1.1. *For any value $0<p<1$ and any parameter (θ, F), there exists a parameter (θ^*, F^*) with $\alpha_0^*=\alpha_0$, $\alpha_1^*=\alpha_1$, $\beta_1^*=\beta_1$ and $\sigma^*=\sigma$ such that*

$$P_{\theta^*, F^*}(E=1)=p,$$

$$\frac{p_{\theta^*}(e, w \mid z)\, dF^*(z)}{P_{\theta^*, F^*}(E=e)}=\frac{p_\theta(e, w \mid z)\, dF(z)}{P_{\theta, F}(E=e)}, \qquad a.s.\ (e, w, z).$$

Furthermore, if the second equation of the display is valid for two arbitrary pairs (θ^*, F^*) and (θ, F) with nondegenerate F^* or F, then $\beta_1^*=\beta_1$.

As shown by Roeder *et al.* (1996), this lemma has several consequences for likelihood inference. Start by noting that the prospective likelihood is the product of the retrospective likelihood and a likelihood of multinomial form. If n_0 and n_1 are the total numbers of controls and cases, respectively, then

$$\text{Pros}(\theta, F)=\text{Retro}(\theta, F)\times P_{\theta, F}(E=0)^{n_0}\, P_{\theta, F}(E=1)^{n_1}.$$

The multinomial likelihood $(1-p)^{n_0}\, p^{n_1}$ is maximized over $0\leqslant p\leqslant 1$ by $p=n_1/n$ for $n=n_0+n_1$. In view of the lemma this value is attained within the class of probabilities $P_{\theta, F}(E=1)$ as (θ, F) ranges over the parameter set. Moreover, this value can be attained meanwhile allowing complete liberty in the value of $\text{Retro}(\theta, F)$. It follows that the maximum likelihood estimator $(\hat\theta, \hat F_n)$ in the prospective model necessarily maximizes both the multinomial likelihood, with $P_{\hat\theta, \hat F_n}(E=1)=n_1/n$, and the retrospective likelihood. Furthermore, the profile likelihoods for β_1 in the prospective and retrospective models are proportional,

$$\text{Prof}(\beta_1):=\sup_{\alpha, \beta_0, \sigma, F}\text{Pros}(\theta, F)=\sup_{\alpha, \beta_0, \sigma, F}\text{Retro}(\theta, F)\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n_1}{n}\right)^{n_1}.$$

Thus, the prospective maximum likelihood estimator for (θ, F) is also a maximum likelihood estimator for the retrospective model. The retrospective maximum likelihood estimators for β_0 and F are not unique, but the prospective and retrospective maximum likelihood estimators for the other parameters are unique and coincide. Roeder *et al.* (1996) use these observations to show that algorithms for computing a maximum likelihood estimator in semiparametric mixture models apply to compute the maximum likelihood estimator in both prospective and retrospective studies.

For the present paper these observations have the consequence that the retrospective maximum likelihood estimator and the retrospective profile likelihood for β_1 can be handled by similar techniques as the prospective procedures. It should be noted, however, that, while algebraically identical,

the distributional theory is different for the two models. We prove the following theorems, where we make the following assumptions concerning the numbers of observations. In the prospective model, we assume that the numbers n_C/n_R converge to a number in $(0, 1)$; in the retrospective model we assume that the numbers n_{C_0}/n_{C_1}, n_0/n_1 and n_{R_0}/n_{R_1} converge to numbers in $(0, 1)$, with remainders smaller than $n^{-1/2}$. (The exact convergence is not necessary, nor is probably the rate $o(n^{-1/2})$, but these assumptions help to keep the statements and proofs easy.)

In the following theorems $\hat{\beta}_{n1}$ is a maximum likelihood estimator for both the retrospective and prospective models. Similar results are valid for the prospective maximum likelihood estimators of the other parameters. (The asymptotic normality of $\hat{F}$ is understood as the asymptotic normality of $\sqrt{n}\int h d(\hat{F}-F)$ as a stochastic process indexed by bounded, Lipschitz functions.)

THEOREM 1.2. *Under both the prospective and the retrospective model, the sequence* $\sqrt{n}(\hat{\beta}_{n1}-\beta_1)$ *is asymptotically normal with mean zero.*

The asymptotic variances in this theorem, which are different in the retrospective and prospective models, are complicated expressions involving the inverse "information operators" of the models. It appears hard to use these expressions directly to construct confidence intervals or carry out tests. The following theorems show that this is not necessary, as we can use the likelihood ratio statistic and the observed (discretized) information, which are directly computable from the likelihood.

THEOREM 1.3. *Under both the prospective and the retrospective model,*

$$2\log\frac{\text{Prof}(\hat{\beta}_{n1})}{\text{Prof}(\beta_1)} \rightsquigarrow \chi_1^2.$$

THEOREM 1.4. *Under both the prospective and the retrospective model, for every random sequence* $\hat{h}_n \to 0$ *such that* $(\sqrt{n}\,\hat{h}_n)^{-1} = O_P(1)$,

$$-2\frac{\log \text{Prof}(\hat{\beta}_{n1}+\hat{h}_n)-\log \text{Prof}(\hat{\beta}_{n1})}{\hat{h}_n^2} \xrightarrow{\text{P}} \sigma^{-2},$$

where σ^2 *is the asymptotic variance of the sequence* $\sqrt{n}(\hat{\beta}_{n1}-\beta_1)$ (*which is different for the two models*).

The remainder of the paper consists of proofs of these results. We have made the assumption that the numbers of reduced and complete observations are of the same order. For simplicity of notation, we shall henceforth even assume that $n_C = n_R$ and denote the common value by n. We pair the

observations with the first member of every pair coming from the reduced sample and the second member of the pair from the complete sample. Thus a typical observation takes the form $(X, Y, Z) = (X, (Y, Z))$, where $X = (D, V)$ is a reduced observation (consisting of a logistic regression D and a linear regression V on an unobserved covariate), and where $(Y, Z) = (E, W, Z)$ is a complete observation, following the model as introduced previously. The total set of observations is denoted by $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$. In the prospective model, this is a random sample from the density

$$(x, y, z) \mapsto \int p_\theta(x \mid s)\, dF(s)\, p_\theta(y \mid z)\, dF(z) =: p_\theta(x \mid F)\, p_\theta(y \mid z)\, dF(z). \tag{1.1}$$

In the retrospective model the observations are not i.i.d., but consist of two independent random samples, the first of n_0 controls and the second of n_1 cases, respectively, from the densities $(\delta \in \{0, 1\})$

$$(v, w, z) \mapsto \frac{\int p_\theta(\delta, v \mid s)\, dF(s)\, p_\theta(\delta, w \mid z)\, dF(z)}{\iint p_\theta(\delta, v \mid s)\, dF(s)\, dv \iint p_\theta(\delta, w \mid z)\, dF(z)\, dw}. \tag{1.2}$$

With these notations, the symbol n corresponds to the total number of paired controls and paired cases, n_0 and n_1. ("Paired" has as a consequence that in the rest of the paper n_0 and n_1 are half the numbers n_0 and n_1 used previously.) In both models we write $\mathbb{P}_n$ for the empirical measure, $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{(X_i, Y_i, Z_i)}$. Thus, in the retrospective model $n = n_0 + n_1$, and the first n_0 of the variables (X_i, Y_i, Z_i) have the form $(0, V_i, 0, W_i, Z_i)$, while the last n_1 observations take the form $(1, V_i, 1, W_i, Z_i)$.

The organization of the paper is as follows. In Section 2 we state the consistency of the maximum likelihood estimators. The proof of this is in Section 10 at the end of the paper. In Section 3 we derive a set of maximum likelihood equations that are the basis of the asymptotic normality proof, and part of the consistency proof. This proof is outlined in Section 4, where details are deferred to Sections 5 and 6. In Section 7 it is shown that the estimator is asymptotically efficient. In Section 8 we prove Theorem 1.3. Finally in Section 9 we discuss the changes that need to be made to deal with different specific models of the general type (1.1).

The proof of Theorem 1.4 is given in Murphy and Van der Vaart (1999), where we develop a general approach to prove this type of result and consider the present model as an example. Most of the technical arguments needed in this approach are also needed for Theorems 1.2 and 1.3, and are carried out in the present paper.

Other authors who have considered the case of unobserved regression coefficients in semiparametric models are Robins *et al.* (1994). They also assume that "surrogate regression variables" replace the unobserved covariates. Rather than modelling the dependence between surrogates and "true" covariates, they suggest the use of estimating equations in order to construct estimators for the unknown parameters of the model. Thus their estimators are not likelihood based, as the estimators proposed by Roeder *et al.* (1996), studied in the present paper. In case the relationship between surrogates and true covariates cannot be modelled realistically, the estimators of Robins *et al.* (1994) are of course preferable.

2. CONSISTENCY

Due to the fact that the likelihood is a product of an "ordinary" and an "empirical" likelihood, none of the standard consistency proofs apply directly. However, the standard methods can be applied in an adapted form after making the following observation. Let $\tilde{F}_n$ be the empirical distribution of $Z_1, ..., Z_n$. Since $\tilde{F}_n$ maximizes the empirical likelihood $F \mapsto \prod_{i=1}^n F\{Z_i\}$ over all probability distributions F, we have that

$$\mathbb{P}_n \log \hat{F}_n\{z\} \leqslant \mathbb{P}_n \log \tilde{F}_n\{z\}.$$

By the definition of the maximum likelihood estimators $(\hat{\theta}_n, \hat{F}_n)$, we have

$$\mathbb{P}_n \log(p_{\hat{\theta}_n}(x \mid \hat{F}_n)\, p_{\hat{\theta}_n}(y \mid z)\, \hat{F}_n\{z\}) \geqslant \mathbb{P}_n \log(p_{\theta_0}(x \mid \tilde{F}_n)\, p_{\theta_0}(y \mid z)\, \tilde{F}_n\{z\}).$$

Together, the two displays yield the inequality

$$\mathbb{P}_n \log(p_{\hat{\theta}_n}(x \mid \hat{F}_n)\, p_{\hat{\theta}_n}(y \mid z)) \geqslant \mathbb{P}_n \log(p_{\theta_0}(x \mid \tilde{F}_n)\, p_{\theta_0}(y \mid z)). \tag{2.1}$$

This is valid for both the prospective and the retrospective maximum likelihood estimators. We can use this inequality as the starting point of a consistency proof, the difference with an "ordinary" consistency proof (such as the one in Kiefer and Wolfowitz (1956)) being the presence of $\tilde{F}_n$ instead of F_0 on the right side. Since $\tilde{F}_n \to F_0$, this causes no problems. We may think of the function $(x, y, z) \mapsto p_\theta(x \mid F)\, p_\theta(y \mid z)$ as the density of a vector (X, Y, Z_0) relative to the dominating measure $d\mu(x)\, d\mu(y) \times dF_0(z)$. (So Z_0 always has the distribution F_0, but X is a regression on an unobservable variable with distribution F.) Unfortunately, the parameter Z is not identifiable from this distribution if the true values of both α_1 and β_1 are 0. This will necessitate to consider this case separately in the following proof, where we use the likelihood equations to obtain the consistency of $\hat{F}$ without unnecessary conditions.

LEMMA 2.1. *In the prospective model* $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$ *and* $\hat{F}_n \xrightarrow{\mathrm{P}} F_0$ *relative to the weak topology under* (θ_0, F_0).

According to Lemma 1.1 there exist parameters (θ_0^*, F_0^*) that yield the same retrospective likelihood as the parameters (θ_0, F_0) and such that $\mathrm{P}_{\theta_0^*, F_0^*}(D=1) = \lim n_1/n$. Our choice of the maximum likelihood estimator for the retrospective model is consistent for this parameter.

LEMMA 2.2. *In the retrospective model* $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0^*$ *and* $\hat{F}_n \xrightarrow{\mathrm{P}} F_0^*$ *relative to the weak topology under* (θ_0, F_0).

3. LIKELIHOOD EQUATIONS

Our proof of asymptotic normality of the sequence of maximum likelihood estimators, and part of the consistency proof, proceeds by showing that any maximum likelihood estimator solves a collection of likelihood equations. Next, the system of equations is linearized and inverted to give the asymptotic distribution of $(\hat{\theta}_n, \hat{F}_n)$, or just $\hat{\beta}_{n1}$. In view of the discussion in the introduction, we can use the likelihood equations for the prospective model for both the prospective and retrospective estimators.

Likelihood equations corresponding to θ can be obtained in the usual manner by partial differentiation of the prospective log likelihood with respect to θ at $\hat{\theta}_n$. This yields

$$\mathbb{P}_n(\dot{\ell}_{\hat{\theta}_n, \hat{F}_n}(x) + \dot{\ell}_{\hat{\theta}_n}(y \mid z)) = 0,$$

where $\dot{\ell}_\theta(y \mid z) = \partial/\partial\theta \log p_\theta(y \mid z)$ is the score function for θ for the conditional density $p_\theta(y \mid z)$, and $\dot{\ell}_{\theta, F}(x)$ is the score function for θ of the mixture density $p_\theta(x \mid F)$, given by

$$\dot{\ell}_{\theta, F}(x) = \frac{\int \dot{\ell}_\theta(x \mid s)\, p_\theta(x \mid s)\, dF(s)}{p_\theta(x \mid F)}.$$

Likelihood equations corresponding to the infinite-dimensional parameter F can be obtained by inserting one-dimensional submodels $t \mapsto \hat{F}_t$ passing through $\hat{F}_n$ in the prospective log likelihood and differentiating with respect to t. In particular, given a bounded, measurable function h and every sufficiently small number $|t|$, we can define a probability measure $\hat{F}_t$ by

$$d\hat{F}_t = \left(1 + t\left(h - \int h\, d\hat{F}\right)\right) d\hat{F}_n.$$

This leads to the likelihood equation

$$\mathbb{P}_n A_{\hat{\theta}_n, \hat{F}_n} h(x, z) - P_{\hat{\theta}_n, \hat{F}_n} A_{\hat{\theta}_n, \hat{F}_n} h = 0,$$

where $P_{\theta, F} g$ is the expectation of $g(X, Y, Z)$ under the prospective model and $A_{\theta, F}$ are the "score operators" given by

$$A_{\theta, F} h(x, z) = B_{\theta, F} h(x) + h(z) = \frac{\int h(s)\, p_\theta(x \mid s)\, dF(s)}{p_\theta(x \mid F)} + h(z).$$

The operators $B_{\theta, F}: L_2(F) \mapsto L_2(p_\theta(\cdot \mid F))$ are the score operators for the mixture part of the model. The Hilbert space adjoint $B^*_{\theta, F}$ of this operator is given by

$$B^*_{\theta, F} g(z) = \int g(x)\, p_\theta(x \mid z)\, d\mu(x).$$

If these operators are viewed as operators between Hilbert spaces, then the images $A_{\theta, F} h$ and $B^*_{\theta, F} g$ are only equivalence classes of functions. Throughout we shall use the versions defined by the preceding equations.

The theory of information in semiparametric models (see Begun *et al.*, 1983, or Van der Vaart, 1991) implies that the "best" asymptotic covariance matrix for estimators of θ in the prospective model is the inverse of the "efficient information matrix"

$$\tilde{I}_{\theta, F} = I_{\theta, F} + J_{\theta, F} - P_{\theta, F}(A_{\theta, F}(I + B^*_{\theta, F} B_{\theta, F})^{-1} B^*_{\theta, F} \dot{\ell}_{\theta, F})\, \dot{\ell}^T_{\theta, F},$$

where $J_{\theta, F}$ is the information matrix for θ for the complete observations and $I_{\theta, F}$ is the information matrix for θ in the reduced observations for known F. (See Section 6 for more details.) According to Lemma 1.1 there exist parameters (θ_0^*, F_0^*) that yield the same retrospective likelihood as the parameters (θ_0, F_0) and such that $\mathrm{P}_{\theta_0^*, F_0^*}(D = 1) = \lim n_1/n$. The asymptotic variance of the maximum likelihood estimator for θ in the retrospective model is given by the same inverse information as above but with (θ_0, F_0) replaced by (θ_0^*, F_0^*). We shall show that $\hat{\theta}$ attains this asymptotic covariance in both models.

4. ASYMPTOTIC NORMALITY

Let $\ell^\infty(H)$ denote the set of all bounded functions $z: H \mapsto \mathbb{R}$ on a given (arbitrary) set H. This is a Banach space with respect to the uniform norm

$$\|z\|_H = \sup_{h \in H} |z(h)|.$$

Let H be the set of all functions $h: \mathscr{Z} \mapsto [-1, 1]$ that are Lipschitz of norm 1: $|h(z_1) - h(z_2)| \leqslant \|z_1 - z_2\|$. This is the unit ball $C_1^1(\mathscr{Z})$ of the space of Lipschitz functions on $\mathscr{Z}$, which we denote by $C^1(\mathscr{Z})$. We identify each probability measure F on $\mathscr{Z}$ with an element of $\ell^\infty(H)$ through $Fh = \int h\, dF$. Then convergence of a sequence F_m viewed as elements of $\ell^\infty(H)$ is identical to weak convergence of the sequence F_m viewed as measures on $\mathscr{Z}$. (See, e.g., Van der Vaart and Wellner, 1996, Theorem 1.12.4.)

Let $W_n = (W_{n1}, W_{n2})$ be the element of $\mathbb{R}^k \times \ell^\infty(H)$ given by

$$W_{n1}(\theta, F) = \mathbb{P}_n(\dot{\ell}_{\theta, F}(x) + \dot{\ell}_\theta(y \mid z)),$$
$$W_{n2}(\theta, F)\, h = \mathbb{P}_n A_{\theta, F} h(x, z) - P_{\theta, F} A_{\theta, F} h.$$

Here k is the dimension of Θ, which is 5 in our example. The map $h \mapsto W_{n2}(\theta, F)\, h$ is indeed uniformly bounded on H, because the conditional expectation operator $B_{\theta, F}$ retains boundedness: $0 \leqslant B_{\theta, F} h \leqslant 1$ for every $h \in H$.

The maximum likelihood estimators $(\hat{\theta}_n, \hat{F}_n)$ are zeros of the maps W_n,

$$W_n(\hat{\theta}_n, \hat{F}_n) \equiv 0.$$

Additionally, W_n can be viewed as a map from the space $\mathbb{L} := \mathbb{R}^k \times \ell^\infty(H)$ into itself with as domain $\mathbb{L}_0$ the product of Θ and the set of probability measures in $\ell^\infty(H)$ under the identification $F \leftrightarrow (F \mapsto Fh)$ introduced previously.

We will need suitable centering functions W. In the prospective model, we simply take W equal to the expectation of W_n under the true distribution $P_0 = P_{\theta_0, F_0}$. This is the element $W = (W_1, W_2)$ of $\mathbb{R}^k \times \ell^\infty(H)$ given by

$$\begin{aligned} W_1(\theta, F) &= P_0(\dot{\ell}_{\theta, F}(x) + \dot{\ell}_\theta(y \mid z)), \\ W_2(\theta, F)\, h &= P_0 A_{\theta, F} h(x, z) - P_{\theta, F} A_{\theta, F} h. \end{aligned} \tag{4.1}$$

With this choice of centering function, we have $W(\theta_0, F_0) = 0$, because the scores $\dot{\ell}_{\theta_0, F_0}$ and $\dot{\ell}_{\theta_0}$ have zero means.

In the retrospective model, the expectations of W_{n1} and W_{n2} are given by

$$\frac{n_0}{n} \frac{P_0(\dot{\ell}_{\theta, F}(0, v)\, 1\{d = 0\} + \dot{\ell}_\theta(0, w \mid z)\, 1\{e = 0\})}{P_0(D = 0)}$$
$$+ \frac{n_1}{n} \frac{P_0(\dot{\ell}_{\theta, F}(1, v)\, 1\{d = 1\} + \dot{\ell}_\theta(1, w \mid z)\, 1\{e = 1\})}{P_0(D = 1)}$$
$$\frac{n_0}{n} \frac{P_0 A_{\theta, F} h(0, v, z)\, 1\{d = e = 0\}}{P_0(D = E = 0)}$$
$$+ \frac{n_1}{n} \frac{P_0 A_{\theta, F} h(1, w, z)\, 1\{d = e = 1\}}{P_0(D = E = 1)} - P_{\theta, F} A_{\theta, F} h.$$

This is equal to the function W defined previously if the fraction of cases n_1/n in the sample is equal to the fraction $P_0(D=1)$ of incidence in the population. Typically, this will not be the case. However, according to Lemma 1.1, for every parameter (θ_0, F_0), there exists a parameter (θ_0^*, F_0^*) that gives the same retrospective model and satisfies $P_{\theta_0^*, F_0^*}(D=1) = \lim n_1/n$. By assumption, the latter limit differs from n_1/n at most by a $o(n^{-1/2})$-term. Therefore, if we use the parameter (θ_0^*, F_0^*) to define P_0 throughout, we can use the same centering function W in the prospective and retrospective models. In the retrospective model this will differ by $o(n^{-1/2})$ from the expectation of the likelihood equations, but this is negligible in the following. The remainder $o(n^{-1/2})$ is uniformly in θ, F, and bounded functions h. Taking the same centering function is technically advantageous, because a considerable part of the effort of the proof goes into proving that the centering function is differentiable with a continuously invertible derivative. For the map given by (4.1) this is carried out in Section 6.

Thus the function W is defined by (4.1) throughout the paper.

The sequence of maximum likelihood estimators will be shown to be asymptotically normal by application of the following proposition. See Van der Vaart and Wellner (1996, Theorem 3.3.1) for a proof. Write parameter and estimator as ψ and $\hat{\psi}_n$, respectively.

PROPOSITION 4.1. *Suppose that W_n and W are random maps and a fixed map from a subset $\mathbb{L}_0$ of a normed space $\mathbb{L}$ into another normed space $\mathbb{M}$ such that*

$$\sqrt{n}(W_n - W)(\psi_0) \rightsquigarrow G, \tag{4.2}$$

$$\|\sqrt{n}(W_n - W)(\hat{\psi}_n) - \sqrt{n}(W_n - W)(\psi_0)\| = o_P^*(1 + \sqrt{n}\,\|\hat{\psi}_n - \psi_0\|), \tag{4.3}$$

$$\|W(\psi) - W(\psi_0) - \dot{W}_0(\psi - \psi_0)\| = o(\|\psi - \psi_0\|), \qquad \psi \to \psi_0, \tag{4.4}$$

for a linear, one-to-one map $\dot{W}_0: \operatorname{lin} \mathbb{L}_0 \subset \mathbb{L} \mapsto \mathbb{M}$ with an inverse $\dot{W}_0^{-1}$ that is continuous on the range of $\dot{W}_0$. If $\hat{\psi}_n \to \psi_0$ in outer probability and $W_n(\hat{\psi}_n) = W(\psi_0) + o_P(n^{-1/2})$, then $\sqrt{n}\,\dot{W}_0(\hat{\psi}_n - \psi_0) = -\sqrt{n}(W_n - W)(\psi_0) + o_P(1)$. Consequently, the sequence $\sqrt{n}(\hat{\psi}_n - \psi_0)$ converges in distribution to $-\dot{W}_0^{-1}G$ (where $\dot{W}_0^{-1}$ is continuously extended to the closure of the range of $\dot{W}_0$).

In the prospective model the process $\sqrt{n}(W_n - W)$ takes the simple form

$$\sqrt{n}(W_n - W)(\theta, F) = (\sqrt{n}(\mathbb{P}_n - P_0)(\dot{\ell}_{\theta,F} + \dot{\ell}_\theta), \sqrt{n}(\mathbb{P}_n - P_0)\, A_{\theta,F}).$$

In the retrospective model this is true as well, up to an $o(1)$-term, but with P_0 replaced by $P_{0*} = P_{\theta_0^*, F_0^*}$. The right side is the empirical process indexed

by the class of functions $\{A_{\theta, F}h : h \in H\} \cup \{\dot{\ell}_{\theta, F} + \dot{\ell}_\theta\}$. Therefore, conditions (4.2)–(4.3) can be ascertained with the help of the theory of empirical processes. Section 5 contains the technical details.

By the multivariate central limit theorem applied to the marginals of the process $\sqrt{n}(W_n - W)$, we see that the limit variable G is a Gaussian process. In view of the tightness of G and the continuity of the linear operator $\dot{W}_0^{-1}$, the variable $\dot{W}_0^{-1}G$ is Gaussian as well. Thus, the sequence $\sqrt{n}(\hat{\beta}_{n1} - \beta_1)$ is asymptotically normal. Its asymptotic variance can, in principle, be computed from a formula for $\dot{W}_0^{-1}$ and the covariance function of G, but there is a better, alternative method to do this, given in Section 7.

5. DONSKER CLASSES

In this section we discuss the verification of conditions (4.2) and (4.3).

5.1. *Prospective Model*

In the prospective model, conditions (4.2) and (4.3) of Proposition 4.1 are certainly satisfied if $\dot{\ell}_{\theta, F} + \dot{\ell}_\theta$ is square-integrable and

$$\{\dot{\ell}_{\theta, F}, \dot{\ell}_\theta, B_{\theta, F}h : h \in H, \|\theta - \theta_0\| < \delta, F \text{ is a distribution function on } \mathscr{X}\}$$

$$\text{is } P_0\text{-Donsker for some } \delta > 0, \tag{5.1}$$

$$H \text{ is } F_0\text{-Donsker}, \tag{5.2}$$

$$\sup_{h \in H} P_0(A_{\theta, F}h - A_0h)^2 \to 0, \qquad \text{as} \quad \theta \to \theta_0 \quad \text{and} \quad F \to F_0, \tag{5.3}$$

$$P_0(\dot{\ell}_{\theta, F} - \dot{\ell}_{\theta_0, F_0} + \dot{\ell}_\theta - \dot{\ell}_{\theta_0})^2 \to 0, \qquad \text{as} \quad \theta \to \theta_0 \quad \text{and} \quad F \to F_0. \tag{5.4}$$

That these conditions are sufficient follows from the properties of Donsker classes. See, for instance, Lemma 3.3.5 in Van der Vaart and Wellner (1996). The set H of bounded Lipschitz functions of norm bounded by 1 possesses bracketing entropy numbers of order $1/\varepsilon$ and is a standard example of a Donsker class. Thus (5.2) is satisfied for our main example. In view of the dominated convergence theorem, condition (5.3) is valid if $B_{\theta, F}h \to B_0h$, pointwise, uniformly in h. This can easily be checked for our example, where H is the class of all Lipschitz functions of norm bounded by 1. Similarly (5.4) can be verified by the dominated convergence theorem. We are left with the verification of (5.1), for which we use the following lemma, proved in Van der Vaart (1996b).

LEMMA 5.1. *Let $\mathcal{X} = \bigcup_{j=1}^{\infty} I_j$ be a partition of $\mathbb{R}$ into bounded, convex sets whose Lebesgue measure is bounded uniformly away from zero and infinity. Let $\mathcal{G}$ be a class of functions $g: \mathcal{X} \mapsto \mathbb{R}$ such that the restrictions $g_{|I_j}$ belong to $C^1_{M_j}(I_j)$ for every j. Then $\mathcal{G}$ is P-Donsker for every probability measure P on $\mathcal{X}$ such that $\sum_{j=1}^{\infty} M_j P^{1/2}(I_j) < \infty$.*

Because one of the arguments of the functions in (5.1) is a 0–1 variable, they are not smooth in the sense of the preceding lemma. However, if the classes of functions obtained by fixing the binary argument to either 0 or 1 are both Donsker when viewed as functions of the remaining argument, then these classes are Donsker. (We state this result formally in Lemma 9.2.) This leads to the following lemma.

LEMMA 5.2. *In the prospective model, conditions* (4.2) *and* (4.3) *are satisfied for $H = C^1_1(\mathcal{Z})$.*

Proof. Straightforward differentiation yields

$$\frac{\partial}{\partial x_i} B_{\theta, F} h(x) = \operatorname{cov}_x \left(h(Z), \frac{\partial}{\partial x_i} \log p_\theta(x \mid Z) \right),$$

where for each x the covariance is computed for the random variable Z having the (conditional) density $z \mapsto p_\theta(x \mid z)\, dF(z)/p_\theta(x \mid F)$. Thus, for a given bounded function h,

$$\left| \frac{\partial}{\partial x_i} B_{\theta, F} h(x) \right| \leqslant \|h\|_\infty \frac{\int \left| \frac{\partial}{\partial x_i} \log p_\theta(x \mid z) \right| p_\theta(x \mid z)\, dF(z)}{\int p_\theta(x \mid z)\, dF(z)}.$$

We now apply Lemma 5.1 to the functions $w \mapsto B_{\theta, F} h(d, w)$, for $d = 0$ and $d = 1$ separately. Since

$$\frac{\partial}{\partial w} \log p_\theta(d, w \mid z) = -\frac{w - \alpha_0 - \alpha_1 z}{\sigma^2},$$

we have that $|\partial/\partial w B_{\theta, F} h(d, w)|$ is bounded by a constant times $\sigma^{-2}(j + |\alpha_0| + |\alpha_1|)$ when $j - 1 \leqslant |w| \leqslant j$. Since the tails (in w) of P_0 are sub-Gaussian, the series $\sum_j j P_0(j - 1 \leqslant |w| \leqslant j)^{1/2}$ converges easily. This proves that the functions $B_{\theta, F} h$ (even with h ranging over a set of uniformly bounded functions that are not necessarily Lipschitz) form a Donsker class.

We can argue similarly for the other functions in (5.1). ∎

5.2. *Retrospective Model*

In the retrospective model, the observations are not i.d.d., but the processes $\sqrt{n}(W_n - W)$ are sums of two empirical processes corresponding to independent random samples. Let P_0^d be the distribution with the density (1.2) evaluated at $(\theta, F) = (\theta_0^*, F_0^*)$, and let F_0^d be the corresponding marginal distribution of Z. Then a set of sufficient conditions for (4.2)–(4.3) is given by, for $d \in \{0, 1\}$,

$$\{\dot{\ell}_{\theta, F}, \dot{\ell}_\theta, B_{\theta, F} h(d, v, z) : h \in H, \|\theta - \theta_0\| < \delta,$$

$$F \text{ is a distribution function on } \mathscr{Z}\} \text{ is } P_0^d\text{-Donsker for some } \delta > 0,$$

$$H \text{ is } F_0^d\text{-Donsker},$$

$$\sup_{h \in H} P_0^d (B_{\theta, F} h(d, v, z) - B_{0*} h(d, v, z))^2 \to 0,$$

$$\text{as } \theta \to \theta_0^* \text{ and } F \to F_0^*,$$

$$P_0^d (\dot{\ell}_{\theta, F}(d, v) - \dot{\ell}_{\theta_0^*, F_0^*}(d, v) + \dot{\ell}_\theta(d, w, z) - \dot{\ell}_{\theta_0^*}(d, w, z))^2 \to 0,$$

$$\text{as } \theta \to \theta_0^* \text{ and } F \to F_0^*.$$

These conditions can be checked by exactly the same methods as for the prospective model. Thus we have the following lemma.

LEMMA 5.3. *In the retrospective model, conditions* (4.2) *and* (4.3) *are satisfied for* $H = C_1^1(\mathscr{Z})$.

6. DIFFERENTIABILITY OF W

A main and non-trivial condition of Proposition 4.1 is the differentiability of the map W and the continuity of the inverse of the derivative. Informally, the derivative $\dot{W} = (\dot{W}_1, \dot{W}_2)$ of the map W at (θ_0, F_0) can be derived as follows. First,

$$\begin{aligned} W_1(\theta, F) - W_1(\theta_0, F_0) &= P_0(\dot{\ell}_{\theta, F} - \dot{\ell}_{\theta_0, F_0}) + P_0(\dot{\ell}_\theta - \dot{\ell}_{\theta_0}) \\ &\approx P_0 \ddot{\ell}_{\theta_0, F_0}(\theta - \theta_0) \\ &\quad + \iint (\dot{\ell}_{\theta_0}(x \mid s) - \dot{\ell}_{\theta_0, F_0}(x))\, p_0(x \mid s)\, d\mu(x)\, d(F - F_0)(s) \\ &\quad + P_0 \ddot{\ell}_{\theta_0}(\theta - \theta_0). \end{aligned}$$

As usual we have that $P_0 \ddot{\ell}_{\theta_0, F_0} = -I_0$ is minus the Fisher information matrix for θ in the reduced observations when $F = F_0$ is known, and the integral $\int \dot{\ell}_{\theta_0, F_0}(x \mid z)\, p_0(x \mid z)\, d\mu(x) = 0$ for every z. Additionally, $J_0 = -P_0 \ddot{\ell}_{\theta_0}$ is the Fisher information matrix for θ for a complete observation. Then the last line can be rewritten as

$$-I_0(\theta - \theta_0) - \int B_0^* \dot{\ell}_{\theta_0, F_0}\, d(F - F_0) - J_0(\theta - \theta_0).$$

The derivative of the second component of W can be obtained in a similar way. Uniformly in h

$$\begin{aligned} W_2(\theta, F)\, h - W_2(\theta_0, F_0)\, h &= -\int A_{\theta, F} h\, d(P_{\theta, F} - P_0) \\ &\approx -\int A_0 h\, d(P_{\theta, F} - P_0) \\ &\approx -\int A_0 h \dot{\ell}^T_{\theta_0, F_0}\, dP_0(\theta - \theta_0) \\ &\quad -\int (I + B_0^* B_0)\, h\, d(F - F_0). \end{aligned}$$

Combination of the preceding displays suggests that the derivative of W at (θ_0, F_0) is given by the map

$$(\theta - \theta_0, F - F_0) \mapsto \begin{pmatrix} \dot{W}_{11} & \dot{W}_{12} \\ \dot{W}_{21} & \dot{W}_{22} \end{pmatrix} \begin{pmatrix} \theta - \theta_0 \\ F - F_0 \end{pmatrix}, \tag{6.1}$$

where

$$\begin{aligned} \dot{W}_{11}(\theta - \theta_0) &= -(I_0 + J_0)(\theta - \theta_0), \\ \dot{W}_{12}(F - F_0) &= -\int B_0^* \dot{\ell}_{\theta_0, F_0}\, d(F - F_0), \\ \dot{W}_{21}(\theta - \theta_0)\, h &= -P_0 A_0 h \dot{\ell}^T_{\theta_0, F_0}(\theta - \theta_0), \\ \dot{W}_{22}(F - F_0)\, h &= -\int (I - B_0^* B_0)\, h\, d(F - F_0). \end{aligned}$$

This derivation is correct, as can be checked by somewhat tedious, but elementary arguments. An intermediate set of sufficient conditions to be verified is given by (9.1)–(9.6).

Proposition 4.1 requires that the derivative operator is continuously invertible on the linear span of the domain of W. For our example this is guaranteed by the following lemma.

LEMMA 6.1. *Let* $H = C_1^1(\mathscr{X})$. *Then the map* $W: \mathbb{R}^k \times \ell^\infty(H) \mapsto \mathbb{R}^k \times \ell^\infty(H)$ *with domain the product of* Θ *and the probability measures on* $\mathscr{X}$ *is differentiable at* (θ_0, F_0) *with derivative* $\dot{W}_0$ *given by* (6.1). *The derivative is one-to-one and has a continuous inverse on the linear span of its range.*

Proof. The continuous invertibility of $\dot{W}$ can be verified by ascertaining the continuous invertibility of the two operators $\dot{W}_{11}$ and $\dot{V} = \dot{W}_{22} - \dot{W}_{21}\dot{W}_{11}^{-1}\dot{W}_{12}$. In that case we have

$$\dot{W}^{-1} = \begin{pmatrix} \dot{W}_{11}^{-1}(\dot{W}_{11} + \dot{W}_{12}\dot{V}^{-1}\dot{W}_{21})\,\dot{W}_{11}^{-1} & -\dot{W}_{11}^{-1}\dot{W}_{12}\dot{V}^{-1} \\ -\dot{V}^{-1}\dot{W}_{21}\dot{W}_{11}^{-1} & \dot{V}^{-1} \end{pmatrix}.$$

The operator $\dot{W}_{11}$ is continuously invertible because the information matrix $I_0 + J_0$ is nonsingular. The second operator has the form

$$\dot{V}(F - F_0)\,h = -\int (I + K)\,h\,d(F - F_0),$$

where the operator K is defined as

$$Kh = B_0^* B_0 h - (P_0 A_0 h \dot{\ell}_{\theta_0, F_0}^T)(I_0 + J_0)^{-1} B_0^* \dot{\ell}_{\theta_0, F_0}. \tag{6.2}$$

The operator $\dot{V}$ is certainly continuously invertible if there exists a positive number ε such that

$$\{(I + K)\,h : h \in H\} \supset \varepsilon H. \tag{6.3}$$

Because $H = C_1^1(\mathscr{X})$ is the unit ball of the Banach space $C^1(\mathscr{X})$, a different way of expressing this condition is that the operator $I + K: C^1(\mathscr{X}) \mapsto C^1(\mathscr{X})$ be continuously invertible. We can verify this by the Fredholm theory for linear operators: if K is a compact operator and $I + K$ is one-to-one, then $I + K$ is continuously invertible (See, for instance, Rudin, 1973, pp. 99–103.)

Thus, we wish to verify that K is compact and that $I + K$ is one-to-one. The operator K is a sum of two operators: $B_0^* B_0$ and a remainder. The "remainder" is a continuous, finite-range operator and hence is compact. The compactness of K follows therefore from the compactness of the information operator $B_0^* B_0$. This can be deduced from the smoothness of the maps $z \mapsto p_\theta(x \mid z)$ for given x. We show this for more general kernels in Lemma 9.4.

That $I+K$ is one-to-one is not immediate, but has a statistical interpretation. It comes down to the efficient information matrix for θ being positive-definite. We discuss this as a separate lemma below. ▌

By definition the efficient information matrix $\tilde{I}_{\theta_0, F_0}$ is the covariance matrix of the projection of the score function $\dot{\ell}_{\theta_0, F_0}(x) + \dot{\ell}_{\theta_0}(y \mid z)$ on the orthocomplement of the range of $A_0: L_2(F_0) \mapsto L_2(P_0)$, which is the score-space for the nuisance parameter F at F_0. Since $A_0^* A_0 = I + B_0^* B_0$ on the mean-zero functions in $L_2(F_0)$, and

$$A_0^*(\dot{\ell}_{\theta_0, F_0} + \dot{\ell}_{\theta_0}) = A_0^* \dot{\ell}_{\theta_0, F_0} = B_0^* \dot{\ell}_{\theta_0, F_0},$$

we have that

$$\tilde{I}_{\theta_0, F_0} = J_0 + I_0 - P_0(A_0(I + B_0^* B_0)^{-1} B_0^* \dot{\ell}_{\theta_0, F_0}) \dot{\ell}^T_{\theta_0, F_0}.$$

This matrix is strictly positive-definite, because the information matrix J_0 for the complete observations is strictly positive-definite, while the second term in $\tilde{I}_{\theta_0, F_0}$ is the efficient information about θ in the reduced observations and hence is nonnegative-definite.

LEMMA 6.2. *The operator $I+K: \ell^\infty(\mathscr{Z}) \to \ell^\infty(\mathscr{Z})$ is one-to-one.*

Proof. If $(I+K)h = 0$, then $F_0(h(I+K)h) = 0$ as well. The latter equation can be rewritten as

$$\begin{aligned} &a_0^T(I_0 + J_0)\, a_0 + a_0^T P_0(A_0 h \dot{\ell}_{\theta_0, F_0}) \\ &\quad + F_0(hB_0^* \dot{\ell}^T_{\theta_0, F_0})\, a_0 + F_0 h^2 + F_0(hB_0^* B_0 h) = 0, \end{aligned} \tag{6.4}$$

for $a_0 = -(I_0 + J_0)^{-1} P_0(A_0 h \dot{\ell}_{\theta_0, F_0})$.

For arbitrary $a \in \mathbb{R}^5$ and $h \in \ell^\infty(\mathscr{Z})$, define $\theta_t = \theta_0 + ta$ and $dF_t = (1 + t(h - F_0 h))\, dF_0$. Then, by direct calculation,

$$\begin{aligned} &\frac{\partial^2}{\partial t^2}_{|\, t=0} P_0 \log p_{\theta_t}(x \mid F_t)\, p_{\theta_t}(y \mid z)\, dF_t(z) \\ &\quad = a^T(I_0 + J_0)\, a + a^T P_0(A_0 h \dot{\ell}_{\theta_0, F_0}) + F_0(hB_0^* \dot{\ell}^T_{\theta_0, F_0})\, a \\ &\qquad + F_0 h^2 + F_0(hB_0^* B_0 h). \end{aligned}$$

By the usual arguments this quantity is minus the information about t in the submodel indexed by (θ_t, F_t). For a given direction $a \neq 0$, this information is minimal for the direction h that is least favorable for estimating the parameter $a^T\theta$. Since the efficient information matrix is nonsingular, this minimal information is positive.

Thus (6.4) implies that $a_0 = 0$. Upon inserting this in the equation $(I+K)\,h = 0$, we find that $(I + B_0^* B_0)\,h = 0$, and upon inserting $a_0 = 0$ in (6.4), we find that $h = 0$ almost surely under F_0. Together this yields that $h = -B_0^* B_0 h = B_0^* 0 = 0$, by the definitions of B_0^* and B_0. ∎

By the same arguments as in the preceding proofs we also have the following lemma, which is used in the consistency proof.

LEMMA 6.3. *For any distribution function F on $\mathscr{X}$, the operator $I + B_0^* B_{\theta_0, F}: C^1(\mathscr{X}) \mapsto C^1(\mathscr{X})$ is continuously invertible.*

7. ASYMPTOTIC LINEARITY AND EFFICIENCY

The asymptotic covariance of the sequence $\sqrt{n}(\hat\theta_n - \theta_0)$ can be computed from the expression (6.1) for $\dot W_0$ and the representation $-\dot W_0^{-1} G$ for the limit distribution of the maximum likelihood estimator. However, it is easier to use an asymptotic representation of $\sqrt{n}(\hat\theta_n - \theta_0)$ as a sum. This is obtained as follows.

By Proposition 4.1 we have that

$$\sqrt{n}\,\dot W_0(\hat\theta_n - \theta_0, \hat F_n - F_0) = -\sqrt{n}(W_n - W)(\theta_0, F_0) + o_P(1).$$

In view of (6.1), this can be rewritten as the system of equations

$$\begin{aligned} -(I_0 + J_0)(\hat\theta_n - \theta_0) - \int B_0^* \dot\ell_{\theta_0, F_0}\, d(\hat F_n - F_0) \\ = -(W_{n1} - W_1)(\theta_0, F_0) + o_P(1/\sqrt{n}), \\ -P_0 A_0 h \dot\ell_{\theta_0, F_0}^T (\hat\theta_n - \theta_0) - \int (I + B_0^* B_0)\, h\, d(\hat F_n - F_0) \\ = -(W_{n2} - W_2)(\theta_0, F_0) + o_P(1/\sqrt{n}). \end{aligned}$$

The $o_P(1/\sqrt{n})$-term in the second line is valid for every $h \in H$ (uniformly in h). If we choose $h = (I + B_0^* B_0)^{-1} B_0^* \dot\ell_{\theta_0, F_0}$, and subtract the first equation from the second, then we arrive at

$$\tilde I_{\theta_0, F_0} \sqrt{n}(\hat\theta_n - \theta_0) = \sqrt{n}(\mathbb{P}_n - P_0)\, \tilde\ell_{\theta_0, F_0} + o_P(1),$$

where

$$\tilde I_{\theta_0, F_0} = I_0 + J_0 - P_0(A_0(I + B_0^* B_0)^{-1} B_0^* \dot\ell_{\theta_0, F_0})\, \dot\ell_{\theta_0, F_0}^T$$

is the efficient information matrix for θ, and $\tilde{\ell}_{\theta_0, F_0}$ is the efficient score function for θ, defined by

$$\tilde{\ell}_{\theta_0, F_0}(x, y, z) = \dot{\ell}_{\theta_0, F_0}(x) + \dot{\ell}_{\theta_0}(y \mid z) - A_0(I + B_0^* B_0)^{-1} B_0^* \dot{\ell}_{\theta_0, F_0}.$$

(The preceding representation for $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is valid for the retrospective model provided we substitute (θ_0^*, F_0^*) for (θ_0, F_0).) The representation shows that $\hat{\theta}_n$ is asymptotically efficient for estimating θ, a conclusion that could also have been reached from general results on the asymptotic efficiency of the maximum likelihood estimator. See Gill and Van der Vaart (1993) and Van der Vaart (1995).

8. LIKELIHOOD RATIO STATISTIC

We establish the asymptotic chi-squared distribution of the likelihood ratio statistic for testing $H_0: \beta_1 = \beta_{10}$ by the method introduced in Murphy and Van der Vaart (1997). This consists of "sandwiching" the likeliood ratio statistic using perturbations of the maximum likelihood estimators $(\hat{\theta}_n, \hat{F}_n)$ and $(\hat{\theta}_0, \hat{F}_0)$ under the full model and the null hypothesis, respectively, in the "least favorable direction."

The latter direction is defined as follows. Partition θ into $\theta = (\beta_1, \theta_2)$, where $\theta_2 = (\alpha_0, \alpha_1, \beta_0, \sigma^2)$, and partition the efficient information matrix $\tilde{I}_0 = \tilde{I}_{\theta_0, F_0}$ accordingly. Define

$$\begin{aligned} a_0^T &= (1, -\tilde{I}_{0,12}(\tilde{I}_{0,22})^{-1}), \\ h_0 &= (I + B_0^* B_0)^{-1} B_0^* \dot{\ell}_{\theta_0, F_0}, \\ dF_t(\theta, F) &= (1 + (\theta - t)^T (h_0 - Fh_0))\, dF, \\ \theta_s(\theta, F) &= (s - \beta_1)\, a_0 + \theta. \end{aligned}$$

The function h_0 is bounded, because the inverse operator $(I + B_0^* B_0)^{-1}$ maps Lipschitz functions into bounded (Lipschitz) functions. Therefore, $F_t(\theta, F)$ has a positive density with respect to F for every sufficiently small $\|\theta - t\|$ and hence defines an element of the parameter set for F. Now define

$$\ell(s, \theta, F)(x, y, z) = \log(p_{\theta_s(\theta, F)}(x \mid G)\, p_{\theta_s(\theta, F)}(y \mid z)\, G\{z\})|_{G = F_{\theta_s(\theta, F)(\theta, F)}}.$$

Then $s \mapsto \exp \ell(s, \theta, F)$ is a one-dimensional submodel of the retrospective model that is least favorable at (θ_0, F_0) in the sense that

$$\frac{\partial}{\partial s}_{|s = \beta_{10}} \ell(s, \theta_0, F_0) = a_0^T \tilde{\ell}_{\theta_0, F_0}.$$

The function on the right is the efficient score function for β_1 in the presence of the nuisance parameter (θ_2, F) at (θ_0, F_0).

The argument by Murphy and Van der Vaart (1997) next uses the inequalities

$$n\mathbb{P}_n(\ell(\hat\beta_{n1}, \hat\theta_0, \hat F_0) - \ell(\beta_{10}, \hat\theta_0, \hat F_0))$$

$$\leqslant \log \frac{\mathrm{Prof}(\hat\beta_{1n})}{\mathrm{Prof}(\beta_{10})} \leqslant n\mathbb{P}_n(\ell(\hat\beta_{n1}, \hat\theta_n, \hat F_n) - \ell(\beta_{10}, \hat\theta_n, \hat F_n)).$$

These are valid trivially by the fact that the estimators $(\hat\theta_n, \hat F_n)$ and $(\hat\theta_0, \hat F_0)$ are maximizers, since $\theta_\theta(\theta, F) = \theta$, $F_\theta(\theta, F) = F$ and $\theta_{\beta_{10}}(\theta, F) = \beta_{10}$. As explained in the introduction, the expression in the middle is the likelihood ratio statistic for both the prospective and retrospective model. The proof proceeds by expanding both extreme sides of this inequality in two-term Taylor expansions in $\hat\beta_{n1} - \beta_{10}$, around β_{10} and $\hat\beta_{n1}$, respectively, leaving the other arguments fixed. Both sides are next shown to be asymptotically equivalent to $\sqrt{n}(\hat\beta_{n1} - \beta_1)^2/(\tilde I^{-1}_{\theta_0, F_0})_{11}$ and hence are asymptotically chi-squared distributed.

As shown in Murphy and Van der Vaart (1997), the only structural condition to carry this through is, with $\dot\ell$ the derivative of ℓ with respect to its first argument,

$$\sqrt{n}\, P_0 \dot\ell(\beta_{10}, \hat\theta_0, \hat F_0) \xrightarrow{\mathrm{P}} 0. \tag{8.1}$$

(In the retrospective model read (θ_0^*, F_0^*) for (θ_0, F_0).) By simple calculus,

$$\dot\ell(\beta_{10}, \hat\theta_0, \hat F_0)(x, y, z) = a_0^T(\dot\ell_{\hat\theta_0, \hat F_0}(x) + \dot\ell_{\hat\theta_0}(y \mid z) - a_0^T A_{\hat\theta_0, \hat F_0}(h_0 - \hat F_0 h_0)(x).$$

It follows that the left side of (8.1) is equal to

$$\begin{aligned}
&\sqrt{n}\, a_0^T(W_1(\hat\theta_0, \hat F_0) - W_2(\hat\theta_0, \hat F_0)\, h_0) \\
&\quad = \sqrt{n}\, a_0^T(\dot W_1(\hat\theta_0 - \theta_0, \hat F_0 - F_0) - \dot W_2(\hat\theta_0 - \theta_0, \hat F_0 - F_0)\, h_0) \\
&\qquad + \sqrt{n}\, o_P(\|\hat\theta_0 - \theta_0\| + \|\hat F_0 - F_0\|_H), \\
&\quad = \sqrt{n}\, o_P(\|\hat\theta_0 - \theta_0\| + \|\hat F_0 - F_0\|_H),
\end{aligned}$$

by the definitions of $\dot W_1 = (\dot W_{11}, \dot W_{12})$, $\dot W_2 = (\dot W_{21}, \dot W_{22})$, h_0 and a_0. The maximum likelihood estimator $(\hat\theta_0, \hat F_0)$ can be shown to be asymptotically normal just as the full maximum likelihood estimator. Condition (8.1) follows.

9. MORE GENERAL MODELS

In this section we indicate the changes that should be made to the preceding discussion if the kernel $p_\theta(x \mid z)$ is different from the one considered by Roeder *et al.* (1996).

First, we note that Lemma 1.1 depends crucially on the case-control indicator E in the basic model being a logistic regression on a function of Z. It does not depend on the model for the distribution of the surrogate W given Z. Hence as long as E is a logistic regression on a function of Z, then Lemma 1.1 remains valid and so do the arguments that connect the retrospective and prospective likelihoods. The Gaussian linear regression of W on Z can be replaced by another model.

Second, the consistency of the maximum likelihood estimators depends on the identifiability of the parameters and regularity conditions. Our proof for the special case is lengthy, because it appears to be necessary to use special properties of our example to deduce consistency without unnatural restrictions. The general ideas of this proof should go through, but different models require work. Consistency proofs always require work.

Third, the derivation of the asymptotic normality of the maximum likelihood estimators should go through along broadly the same lines in some generality. However, this derivation requires a number of steps and each step may need to be adapted. We have no hope to write up a single theorem that is general enough to cover most cases of interest.

We discuss this in more detail. The likelihood equations, derived in Section 3 are written in general notation and need not be adapted. We still would obtain the asymptotic normality of the maximum likelihood estimators as outlined in Section 4, by application of Proposition 4.1, but the normed spaces involved in this proposition may need to be chosen differently. More specifically, we follow Section 4 as is, except that we do not immediately fix the set of functions H as the set of all Lipschitz functions of norm bounded by 1. Other potentially useful choices are the unit balls in the set of functions of bounded variation, or in one of the Hölder classes $C^\alpha(\mathscr{Z})$. These are the spaces of functions $h: \mathscr{Z} \to \mathbb{R}$ that have continuous (partial) derivatives up to order β for β the largest integer less than or equal than α and whose partial derivatives of order β are uniformly Lipschitz of order $\alpha - \beta$. Choosing a unit ball relative to some norm is potentially convenient to push through the argument for continuous invertibility in Section 6. The particular choice of the Lipschitz norm made in Section 4 is motivated by the fact that $\mathscr{Z}$ is one-dimensional and the kernel $p_\theta(x \mid z)$ smooth in z.

9.1. *Donsker Classes*

For the verification of conditions (4.2) and (4.3) of Proposition 4.1 we may again check the validity of (5.1)–(5.4). Here (5.3)–(5.4) remain as

primitive conditions, to be checked for particular examples, but should not cause trouble. Conditions (5.1)–(5.2) are more involved. We should keep them in mind when choosing the class of functions H indexing the likelihood equations in Section 4. If we choose this set too big, or of the wrong type, then (5.1)–(5.2) will fail.

There is a large literature on empirical processes, and this is not easily summarized. The most recent reviews are Dudley (1984), Pollard (1984, 1990), Giné and Zinn (1986), and Van der Vaart and Wellner (1996).

Condition (5.2) is clear in its demand: we can just pick one of the known Donsker classes for our indexing set H (and then must move on to see whether (5.1) is satisfied and whether this H makes the map W differentiable with continuous inverse). If we choose a unit ball in a Hölder space $C^\alpha(\mathscr{X})$, then we must choose $\alpha > d/2$, for d the dimension of $\mathscr{X}$, for otherwise H will not be Donsker. So our earlier choice $H = C_1^1(\mathscr{X})$ can only work if $\mathscr{X}$ is one-dimensional, which is a severe limitation.

To satisfy (5.1) one possibility is to use the fact that classes of smooth functions are Donsker classes. If the kernels $x \mapsto p_\theta(x \mid z)$ are smooth functions, as is the case in many examples, then the functions $x \mapsto B_{\theta, F} h(x)$ are smooth also.

If the variable $x = (d, v)$ is partitioned into a discrete variable d and a continuous variable v, as it is in the logistic regression case, then "smoothness in d" does not make sense. However, discrete variables can be handled in a trivial way, and therefore we may focus our attention on the smooth part of x. This follows from Lemma 9.2 below.

An appropriate lemma about Donsker classes of smooth functions on possibly unbounded subsets of $\mathbb{R}^d$ is as follows (cf. Van der Vaart (1996b)).

LEMMA 9.1. *Let $\mathscr{X} = \bigcup_{j=1}^\infty I_j$ be a partition of $\mathbb{R}^d$ into bounded, convex sets whose Lebesgue measure is bounded uniformly away from zero and infinity. Let $\mathscr{G}$ be a class of functions $g: \mathscr{X} \mapsto \mathbb{R}$ such that the restrictions $g_{|I_j}$ belong to $C^\alpha_{M_j}(I_j)$ for every j and some fixed $\alpha > d/2$. Then $\mathscr{G}$ is P-Donsker for every probability measure P on $\mathscr{X}$ such that $\sum_{j=1}^\infty M_j P^{1/2}(I_j) < \infty$.*

We can establish bounds on the Hölder norms of order 1 of the functions $B_{\theta, F} h(x)$ by the same method as in the proof of Lemma 5.2. However, if $\mathscr{X}$ is a subset of a higher-dimensional Euclidean space, then the preceding lemma requires consideration of higher-order derivatives. For instance, in dimension two any Lipschitz condition on the first order partial derivatives suffices ($\alpha > 1$), while in dimension three we need a Lipschitz condition of order $> 1/2$ on these derivatives ($\alpha > 3/2$). Straightforward calculations show that

$$\frac{\partial^2}{\partial x_i \partial x_j} B_{\theta, F} h(x) = \operatorname{cov}_x \left(h(Z), \frac{\partial^2}{\partial x_i \partial x_j} \log p_\theta(x \mid Z) \right)$$
$$- \operatorname{cov}_x \left(h(Z), \frac{\partial}{\partial x_i} \log p_\theta(x \mid Z) \right) E_x \frac{\partial}{\partial x_j} \log p_\theta(x \mid Z)$$
$$- \operatorname{cov}_x \left(h(Z), \frac{\partial}{\partial x_j} \log p_\theta(x \mid Z) \right) E_x \frac{\partial}{\partial x_i} \log p_\theta(x \mid Z).$$

This expression can be bounded as before. For $\mathcal{X}$ of dimension four and five, we must also consider the third order derivatives, etcetera. Then this method will lead to increasingly stringent conditions on the kernel $x \mapsto p_\theta(x \mid z)$.

Finally we note that imposing smoothness conditions is only one method to verify the Donsker condition. For instance, in a related problem Van der Vaart (1994) also discusses examples with discontinuous kernels.

LEMMA 9.2. *Let $\mathcal{F}$ be a class of measurable functions $f: \mathbb{D} \times \mathbb{W} \mapsto \mathbb{R}$ on a product of a finite set and an arbitrary measurable space $(\mathbb{W}, \mathcal{W})$. Let P be a probability measure on $\mathbb{D} \times \mathbb{W}$ and let P_W be its marginal on $\mathbb{W}$. For every $d \in \mathbb{D}$, let $\mathcal{F}_d$ be the set of functions $w \mapsto f(d, w)$ as f ranges over $\mathcal{F}$. If every class $\mathcal{F}_d$ is P_W-Donsker with $\sup_f |Pf(d, W)| < \infty$ for every d, then $\mathcal{F}$ is P-Donsker.*

Proof. The empirical process indexed by $\mathcal{F}$ of a random sample $(D_1, W_n), ..., (D_n, W_n)$ from P can be written as

$$\sum_d \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(f(d, W_i) 1\{D_i = d\} - Pf(d, W_1) 1\{D_1 = d\} \right).$$

Under the condition of the lemma, the class $\mathcal{G}$ of functions $(w, d) \mapsto f(d_0, w) 1\{d = d_0\}$ is P-Donsker for every fixed d_0. To see this, we first note that the class of functions $\mathcal{F}_{d_0}$ viewed as functions on $\mathbb{D} \times \mathbb{W}$ (that depend on the second coordinate of (d, w) only) is P-Donsker. (Measurability problems do not occur for the coordinate projection $(d_1, w_1, ..., d_n, w_n) \mapsto (w_1, ..., w_n)$ is perfect in view of the finiteness of $\mathbb{D}$.) Next, the claim follows from Example 2.10.10 in Van der Vaart and Wellner (1996), because the class $\mathcal{G}$ is of the form $\mathcal{F}_d g$ with g bounded. Thus, the processes in the sum in the preceding display converge in distribution to a tight limit in $\ell^\infty(\mathcal{F})$, for every d. So does their sum, by marginal convergence and the continuous mapping theorem. ∎

9.2. *Differentiability*

The other main part of the verification of the conditions of Proposition 4.1 is a proof that the centering function W is differentiable with continuously invertible derivative.

The map W is differentiable at (θ_0, F_0) with the derivative given by (6.1) as before under the following conditions. The first component W_1 is differentiable under the conditions, as $\theta \to \theta_0$ and $F \to F_0$,

$$P_0 \|\dot{\ell}_{\theta, F} - \dot{\ell}_{\theta_0, F} - \ddot{\ell}_{\theta_0, F_0}(\theta - \theta_0)\| = o(\|\theta - \theta_0\|), \tag{9.1}$$

$$\int \left[\int (\dot{\ell}_{\theta_0, F}(x) - \dot{\ell}_{\theta_0, F_0}(x))\, p_{\theta_0}(x \mid z)\, d\mu(x) \right] d(F - F_0) = o(\|F - F_0\|_H) \tag{9.2}$$

$$P_0 \|\dot{\ell}_{\theta} - \dot{\ell}_{\theta_0} - \ddot{\ell}_{\theta_0}(\theta - \theta_0)\| = o(\|\theta - \theta_0\|). \tag{9.3}$$

The second component W_2 is differentiable under the conditions

$$\int |p_\theta(x \mid F_0) - p_{\theta_0}(x \mid F_0) - \dot{\ell}_{\theta_0, F_0}(x)\, p_{\theta_0}(x \mid F_0)(\theta - \theta_0)|\, d\mu(x) = o(\|\theta - \theta_0\|), \tag{9.4}$$

$$\sup_{h \in H} P_0(B_{\theta, F} h - B_0 h)^2 = o(1), \tag{9.5}$$

$$\sup_{h \in H} \left| \int (B^*_{\theta, F} B_{\theta, F} h - B_0^* B_0 h)\, d(F - F_0) \right| = o(\|F - F_0\|_H). \tag{9.6}$$

These conditions are reasonable, but remain as primitive conditions to be checked for particular examples. In the case that $H = C_1^\beta(Z)$, condition (9.2) can be checked by showing that the term in square brackets converges to zero in the $C^\beta(Z)$-norm as $\theta \to \theta_0$ and $F \to F_0$. Condition (9.6) follows if $B^*_{\theta, F} B_{\theta, F} \to B_0^* B_0$ in operator norm as operators from $C^\beta(\mathscr{Z})$ into itself.

Given the structure of the information operator as a sum of the identity and another operator, it is tempting to use Fredholm theory to verify its continuous invertibility, as we did in Section 6. The approach in Section 6 can be summarized as follows.

LEMMA 9.3. *Suppose that H is the unit ball of a Banach space $\mathbb{B}$ of functions, contained in $\ell^\infty(\mathscr{Z})$. Let $\dot{W}_0$ be given by* (6.1). *Then $\dot{W}_0$ is continuously invertible if $K: \mathbb{B} \mapsto \mathbb{B}$ is compact and the Fisher information matrix J_0 for θ at θ_0 in a complete observation is nonsingular.*

Proof. By Lemma 6.2 the operator $I + K: \mathbb{B} \mapsto \mathbb{B}$ is one-to-one. Therefore, by the Fredholm theory (cf. Rudin, 1973, pp. 99–103) the operator $I + K: \mathbb{B} \mapsto \mathbb{B}$ is continuously invertible if K is compact. This implies that (6.3) is satisfied. The conclusion follows as in the proof of Lemma 6.1. ∎

Whether the operator $K: \mathbb{B} \mapsto \mathbb{B}$ is compact depends on the Banach space $\mathbb{B}$ and its norm. The following lemma gives easily verifiable conditions for compactness relative to Hölder norms. Because the second part of

K is a finite-range operator, this is certainly compact. Therefore, we may concentrate on the first part of K, the operator $B_0^* B_0$.

LEMMA 9.4. *Let $\mathscr{Z}$ be a bounded convex subset of $\mathbb{R}^d$ and assume that the maps $z \mapsto p_0(x \mid z)$ are continuously differentiable for each x with partial derivatives $\partial/\partial z_i \, p_{\theta_0}(x \mid z)$ satisfying, for all z, z' in $\mathscr{Z}$ and fixed constants D and $\alpha > 0$,*

$$\int \left| \frac{\partial}{\partial z_i} p_0(x \mid z) - \frac{\partial}{\partial z_i} p_0(x \mid z') \right| d\mu(x) \leqslant D \, \|z - z'\|^{\alpha},$$

$$\int \left| \frac{\partial}{\partial z_i} p_0(x \mid z) \right| d\mu(x) \leqslant D.$$

Then the range of the operator B_0^* restricted to the domain $\ell^\infty(\mathscr{X})$ is contained in $C^{1+\alpha}(\mathscr{Z})$. Additionally $B_0^*: \ell^\infty(\mathscr{Z}) \mapsto C^{1+\beta}(\mathscr{Z})$ is compact for every $\beta < \alpha$. Consequently, the operator $K: C^\gamma(\mathscr{Z}) \mapsto C^\gamma(\mathscr{Z})$ is compact for every $0 \leqslant \gamma < 1 + \alpha$.

Proof. It follows from the Lipschitz condition on the partial derivatives that $B_0^* g(z)$ is differentiable for every bounded function $g: \mathscr{X} \mapsto \mathbb{R}$ and its partial derivatives can be found by differentiating under the integral sign:

$$\frac{\partial}{\partial z_i} B_0^* g(z) = \int g(x) \frac{\partial}{\partial z_i} p_0(x \mid z) \, d\mu(x).$$

The two conditions of the lemma imply that this function has Lipschitz norm of order α bounded by $K \|g\|_\infty$. Let g_n be a uniformly bounded sequence in $\ell^\infty(\mathscr{X})$. Then the partial derivatives of the sequence $B_0^* g_n$ are uniformly bounded and have uniformly bounded Lipschitz norms of order α. Since $\mathscr{Z}$ is totally bounded, it follows by a strengthening of the Arzela–Ascoli theorem that the sequences of partial derivatives are precompact with respect to the Lipschitz norm of order β for every $\beta < \alpha$. Thus there exists a subsequence along which the partial derivatives converge in the Lipschitz norm of order β. By the Arzela–Ascoli theorem there exists a further subsequence such that the functions $B_0^* g_n(z)$ converge uniformly to a limit. If both a sequence of functions itself and their continuous partial derivatives converge uniformly to limits, then the limit of the functions must have the limits of the sequences of partial derivatives as its partial derivatives. Conclude that $B_0^* g_n$ converges in the $\|\cdot\|_{1+\beta}$-norm.

If the operator $B_0^*: \ell^\infty(\mathscr{X}) \mapsto C^\gamma(\mathscr{Z})$ is compact, then for any distribution function, F on $\mathscr{Z}$, the operator $B_0^* B_{\theta_0, F}$ is certainly compact as an operator from $C^\gamma(\mathscr{Z})$ into itself. The second part of K is always compact, because it has a finite-dimensional range. ∎

10. PROOF OF LEMMAS 2.1 AND 2.2

We shall first give the proof of Lemma 2.1 under the additional assumption that the likelihood is maximized with respect to θ over a compact subset of the natural parameter set Θ. At the end of the proof we indicate how this assumption can be omitted by a compactification argument. We abbreviate P_{θ_0, F_0} by P_0.

In the model for (X, Y, Z_0) as described previously, the parameter θ is identifiable. Indeed, since F_0 is assumed to be nondegenerate, θ is identifiable from $p_\theta(y \mid z)$. Conclude that

$$P_0 \log p_\theta(x \mid F)\, p_\theta(y \mid z) < P_0 \log p_{\theta_0}(x \mid F_0)\, p_{\theta_0}(y \mid z), \qquad \theta \neq \theta_0.$$

Write $\psi = (\theta, F)$ and define functions

$$m_{\psi, F_2}(x, y, z) = \log \frac{p_\theta(x \mid F)\, p_\theta(y \mid z)}{p_{\theta_0}(x \mid F_2)\, p_{\theta_0}(y \mid z)}.$$

Then, by the arguments preceding the lemma, and the identifiability of θ,

$$\mathbb{P}_n m_{\hat\psi, \tilde F} \geqslant 0,$$

$$P_0 m_{\psi, F_0} < 0, \qquad \text{every} \quad \theta \neq \theta_0.$$

The functions $(\psi, F_2) \mapsto m_{\psi, F_2}(x, y, z)$ are continuous for every (x, y, z). (Continuity in F and F_2 is with respect to the weak topology.) Furthermore, for every ψ and sufficiently small neighbourhoods U of ψ and V_ψ of F_0,

$$P_0 \sup_{\psi_1 \in U, F_2 \in V} m_{\psi_1, F_2} < \infty. \tag{10.1}$$

As in the consistency proof of Wald (1949), this allows, by invoking the monotone convergence theorem along sequences of neighbourhoods shrinking to ψ and F_0, to construct for every $\psi = (\theta, F)$ such that $\theta \neq \theta_0$ a neighbourhood U_ψ and a neighbourhood V of F_0 such that

$$P_0 \sup_{\psi_1 \in U_\psi, F_2 \in V_\psi} m_{\psi_1, F_2} < 0. \tag{10.2}$$

Fix $\varepsilon > 0$. The compact set $\{\psi: \|\theta - \theta_0\| \geqslant \varepsilon\}$ is covered by finitely many of the neighbourhoods U_{ψ_j}. If $\hat\psi \in U_{\psi_j}$ and $\tilde F \in V = \cap V_{\psi_j}$, then the supremum of $\mathbb{P}_n m_{\psi, F_2}$ over $\psi \in U_{\psi_j}$ and $F_2 \in V_{\psi_j}$ is nonnegative. Thus,

$$\mathrm{P}_0(\|\hat\theta - \theta_0\| \geqslant \varepsilon) \leqslant \sum_j \mathrm{P}_0(\mathbb{P}_n \sup_{\substack{\psi \in U_{\psi_j} \\ F_2 \in V_{\psi_j}}} m_{\psi, F_2} \geqslant 0) + \mathrm{P}_0(\tilde F \notin V). \tag{10.3}$$

Each of the probabilities in the sum on the right converges to zero, since the variables inside the probabilities converge to negative constants, by the law of large numbers. The last term on the right converges to zero by the consistency of the empirical distribution.

This concludes the proof of consistency of $\hat{\theta}$ under the assumption that θ is restricted to a compact set. To remove this unnecessary assumption, we shall compactify Θ to its one-point compactification $\bar{\Theta} = \Theta \cup \{\infty\}$. It appears that there is no useful extension of the functions $m_{\psi, F}(x)$ to this compactification. For this reason, following an idea in the last section of Kiefer and Wolfowitz (1956), we first group the n observations into sets of 3 observations plus, if necessary, a set of 4 or 5 observations at the end. The grouped observation $(X_1, Y_1, Z_1, X_2, Y_2, Z_2, X_3, Y_3, Z_3)$ has the density

$$p_{\theta, F}(x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3) = \prod_{i=1}^{3} p_\theta(x_i \mid F)\, p_\theta(y_i \mid z_i).$$

Under the condition that the true parameter F_0 is continuous, we shall show that this density can be continuously extended from Θ to $\bar{\Theta}$ by defining it to be 0 at $\theta = \infty$. (The case of general F_0 needs an additional argument, which is given at the end of the proof.) Furthermore, the analog of (10.2) is valid for this extension, i.e., for every ψ (from $\bar{\Theta}$ times the parameter set for F) and sufficiently small neighbourhoods U_ψ of ψ and V_ψ of F_0

$$P_0 \sup_{\psi_1 \in U_\psi, F_2 \in V_\psi} \sum_{i=1}^{3} m_{3, \psi_1, F_2}(x_i, y_i, z_i) < \infty. \tag{10.4}$$

For a final group of 4 or 5 observations similar results are valid.

The true parameter θ_0 is trivially identifiable with respect to the additional point ∞, and therefore identifiable in the compactified model, relative to the grouped observations.

The consistency proof now follows the same lines as the proof given previously under the assumption that Θ is compact, except that we use the law of large numbers on the approximately $n/3$ variables

$$\sup_{\psi_1 \in U_{\psi_j}, F_2 \in V} \sum_{k=1}^{3} m_{\psi_1, F_2}(X_{3i+k}, Y_{3i+k}, Z_{3i+k}), \qquad i = 0, 1, 2, \ldots,$$

and the remaining block of 4 or 5 variables, if there is one, rather than on the n variables $\sup m_{\psi_1, F_2}(X_i, Y_i, Z_i)$.

To verify (10.4) and the continuity of the extension of $p_{\theta, F}$ to $\bar{\Theta}$, we first note that $p_{\theta, F}(x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3)$

$$\leqslant \left(\frac{1}{\sigma}\right)^3 \phi(0)^3 \prod_{i=1}^{3} \left(\frac{1}{1+e^{-\beta_0-\beta_1 z_i}}\right)^{d_i} \left(\frac{e^{-\beta_0-\beta_1 z_i}}{1+e^{-\beta_0-\beta_1 z_i}}\right)^{1-d_i}$$

$$\times \left(\frac{1}{\sigma}\right)^3 e^{-1/2\sigma^2 \sum_{i=1}^{3} (w_i-\alpha_0-\alpha_1 z_i)^2}$$

$$\leqslant \left(\frac{1}{\sigma}\right)^6 \phi(0)^3 e^{-(1/2)\, SS_r/\sigma^2},$$

for $SS_r = \inf_{\alpha_0, \alpha_1} \sum_{i=1}^{3} (w_i-\alpha_0-\alpha_1 z_i)^2$. The vectors (w_1, w_2, w_3) and (z_1, z_2, z_3) are linearly independent. Suppose that (z_1, z_2, z_3) are not all equal. If the true parameter F_0 is continuous, then this is true for almost all realizations of the corresponding observations. Then, as $\sigma \to 0$, and trivially when $\sigma \to \infty$, since $SS_r > 0$,

$$\sup_F p_{\theta, F}(x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3) \to 0. \tag{10.5}$$

Second, we have that for $z_i \neq z_j$,

$$\left\| \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \right\| \leqslant \left\| \begin{pmatrix} 1 & z_i \\ 1 & z_j \end{pmatrix}^{-1} \right\| \left\| \begin{pmatrix} \alpha_0+\alpha_1 z_i \\ \alpha_0+\alpha_1 z_j \end{pmatrix} \right\|.$$

This implies that at least one of $|\alpha_0+\alpha_1 z_i|$ or $|\alpha_0+\alpha_1 z_j|$ converges to ∞ if $(\alpha_0, \alpha_1) \to \infty$. In that case (10.5) is valid. Third, by a similar argument, Eq. (10.5) is also true when $(\beta_0, \beta_1) \to \infty$. Each time as one of σ, α or β becomes extreme, the convergence in (10.5) is uniform in the remaining parameters. We conclude that the map $(\theta, F) \mapsto p_{\theta, F}(x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3)$ is continuous, for almost all of its arguments.

Conditionally on z_1, z_2, z_3, the variable SS_r possesses, up to a constant, a chi-square distribution with 1 or 2 degrees of freedom. It follows that $-P_0 \log SS_r < \infty$. Consequently,

$$P_0 \sup_\sigma \log \left(\frac{1}{\sigma}\right)^6 \phi(0)^3 e^{-SS_r/\sigma^2} < \infty.$$

We also have that

$$P_0 \sup_{F_2} -\log p_{\theta_0}(x \mid F_2)\, p_{\theta_0}(y \mid z) < \infty.$$

This concludes the proof of (10.4).

This concludes the proof of consistency of $\hat{\theta}$ under the assumption that the true F_0 is continuous. If F_0 contains a discrete component, then the set $B = \{(x_1, y_1, ..., z_3) : z_1 = z_2 = z_3\}$ has positive measure under P_0, and the preceding argument must be adapted. In this case the extension $p_{\theta, F}(x_1, y_1, ..., z_3)$ may be discontinuous at $\theta = \infty$ when its argument is in B. Actually, this continuity is only used in an intermediate step of the preceding proof and is not necessary for the proof as a whole. First, note that since F_0 is by assumption nondegenerate, the probability of B is less than 1. For $(x_1, y_1, ..., z_3)$ in the complement of B, the functions $p_{\theta, F}(x_1, y_1, ..., z_3)$ are continuous at $\theta = \infty$. Furthermore, the preceding argument shows that (10.4) is valid in the stronger form

$$M := \sup_z \mathrm{E}_0 \left(\sup_{\psi_1, F_2} \sum_{i=1}^{3} m_{\psi_1, F_2}(X_i, Y_i, Z_i) \mid (Z_1, Z_2, Z_3) = z \right) < \infty.$$

Then, for U_m and V_m sequences of decreasing neighbourhoods of $\psi = (\infty, F_1)$ and F_0, respectively, we have, as $m \to \infty$,

$$\begin{aligned} P_0 & \sup_{\psi_1 \in U_m, F_2 \in V_m} \sum_{i=1}^{3} m_{\psi_1, F_2}(x_i, y_i, z_i) \\ & = P_0 \sup_{\psi_1 \in U_m, F_2 \in V_m} \sum_{i=1}^{3} m_{\psi_1, F_2}(x_i, y_i, z_i)\, 1_B \\ & \quad + P_0 \sup_{\psi_1 \in U_m, F_2 \in V_m} \sum_{i=1}^{3} m_{\psi_1, F_2}(x_i, y_i, z_i)\, 1_{B^c} \\ & \leqslant M + P_0 - \infty 1_{B^c} + o(1) = -\infty. \end{aligned}$$

Thus, there exist neighbourhoods U_ψ and V_ψ of $\psi = (\infty, F_1)$ and F_0 such that the left side is negative. This suffices for the proof as given before.

Finally, we prove the consistency of $\hat{F}$. (Actually, the preceding proof yields the consistency of both $\hat{\theta}$ and $\hat{F}$ provided that (θ_0, F_0) is identifiable from the distribution of (X, Y, Z_0). However, this would necessitate the unnecessary condition that the true value of α_1 or β_1 is nonzero. So we give a separate proof.) In view of the likelihood equations, we have, for every bounded function h,

$$0 = (\mathbb{P}_n - P_0) A_{\hat{\theta}, \hat{F}} h + P_0 (A_{\hat{\theta}, \hat{F}} h - A_{\theta_0, \hat{F}} h) + (P_0 - P_{\theta_0, \hat{F}}) A_{\theta_0, \hat{F}} h.$$

(Note that $P_{\theta, F} A_{\theta, F} h = 2Fh$ for all (θ, F), independent of θ.) The first and the second of the three terms on the right converge to zero in probability

uniformly in h ranging over the class $H = C_1^1(\mathscr{Z})$ of Lipschitz functions $h: \mathscr{Z} \mapsto \mathbb{R}$ with Lipschitz constant 1. For the first this follows, because the class of functions $A_{\theta, F}h$ is Glivenko–Cantelli, when θ ranges over a neighbourhood of θ_0, F ranges over all probability distributions on $\mathscr{Z}$, and h ranges over $C_1^1(\mathscr{Z})$. In Eq. (5.1) ahead we even verify that this class of functions is Donsker. The absolute value of the second term can be bounded by $\sup_{h \in H, F} |P_0(B_{\theta, F}h - B_{\theta_0, F}h)|$. This converges to zero by the dominated convergence theorem, if $|B_{\theta, F}h - B_{\theta_0, F}h| \to 0$, pointwise, uniformly in h and F. Since H and the class of functions $\{z \mapsto p_\theta(x \mid z) : \|\theta - \theta_0\| < \varepsilon\}$ are uniformly bounded and equicontinuous, this is the case.

We conclude that the third term on the right side of the preceding display, which can be rewritten as $\psi(\hat{F})$ for $\psi(F)\, h = \int (I + B_0^* B_{\theta_0, F})\, h\, d(F_0 - F)$, converges to zero in probability, uniformly in $h \in H$. We shall show that this implies that $\hat{F} \xrightarrow{\mathrm{P}} F_0$.

First, the map $F \mapsto \psi(F)$ is continuous in the sense that $\psi(F)\, h \to \psi(F_1)\, h$ uniformly in $h \in H$ as $F \to F_1$ weakly. To see this, note that

$$|\psi(F) - \psi(F_1)| \leqslant \left| \int (I + B_0^* B_{\theta_0, F})\, h\, d(F - F_1) \right| + \left| \int (B_0^* B_{\theta_0, F} - B_0^* B_{\theta_0, F_1})\, h\, dF_1 \right|.$$

By Lemma 9.4, the class of functions $\{(I + B_0^* B_{\theta_0, F})\, h : h \in H, F, \text{ a distribution function}\}$ is uniformly bounded and equicontinuous. Therefore, the first term on the right of the preceding display converges to zero, uniformly in h. We may use the dominated convergence theorem to show that the second term converges to zero uniformly in $h \in H$ as well.

Second, by Lemma 6.3 we can write every bounded Lipschitz function h in the form $h = (I + B_0^* B_{\theta_0, F})\, \tilde{h}$ for some bounded Lipschitz function $\tilde{h}$. Thus $\psi(F) = 0$ implies that $\int h\, d(F - F_0) = 0$ for every bounded Lipschitz function h and hence $F = F_0$.

Now the continuity of ψ, the uniqueness of its zero F_0, and the compactness of the set of distribution functions for the weak topology, show that $\psi(\hat{F}) \xrightarrow{\mathrm{P}} 0$ implies that $\hat{F} \xrightarrow{\mathrm{P}} F_0$. This concludes the proof of Lemma 2.1.

In order to prove Lemma 2.2 we may assume that the observations are sampled according to the parameter (θ_0^*, F_0^*), rather than (θ_0, F_0), since this gives the same retrospective likelihood. Next, the proof follows the same steps, with minor changes, where we replace (θ_0, F_0) by (θ_0^*, F_0^*) throughout. A key identity is that, for any h and g,

$$
\begin{aligned}
&\mathrm{E}_{\theta_0^*, F_0^*} \mathbb{P}_n(h(x) + g(y, z)) \\
&\quad = \frac{n_0/n}{P_{0*}(D=0)} P_{0*}(h(0, v)\, 1\{d=0\} + g(0, w, z)\, 1\{e=0\}) \\
&\qquad + \frac{n_1/n}{P_{0*}(D=1)} P_{0*}(h(1, v)\, 1\{d=1\} + g(1, w, z)\, 1\{e=1\}) \\
&\quad \to P_{0*}(h(x) + g(y, z)).
\end{aligned}
$$

Here the expectation on the left is relative to the retrospective model, while the expectation on the right is for (X, Y, Z) a typical observation from the prospective model. By the law of large numbers it follows that, for integrable h and g, $\mathbb{P}_n(h(x) + g(y, z)) \xrightarrow{\mathrm{P}} P_{0*}(h(x) + g(y, z))$. This convergence is uniform over Glivenko–Cantelli classes of functions $v \mapsto h(\delta, v)$ and $(w, z) \mapsto g(\delta, w, z)$ (for $\delta = 0$ and $\delta = 1$).

As a first application of this, we have that the empirical distribution $\tilde{F}_n$ of $Z_1, \ldots, Z_n$ converges in probability to F_0^*. Thus, the second term on the right in (10.3) still converges to zero.

The preceding law of large numbers also applies to the functions m_{ψ, F_2} that are defined in the proof of Lemma 2.1, but not necessarily to the functions $\sup_{\psi \in U, F_2 \in V} m_{\psi, F_2}$, because the latter lack the structure of a sum of a function of x and a function of (y, x). To cope with this difficulty, we may replace these suprema by the functions

$$
\sup_{\theta \in U_1, \theta' \in U_1, F \in U_2, F_2 \in V} \log \frac{p_\theta(x \mid F)\, p_{\theta'}(y \mid z)}{p_{\theta_0}(x \mid F_2)\, p_{\theta_0}(y \mid z)}.
$$

The analogon of (10.2) is valid for these functions: for every $\psi = (\theta, F)$ such that $\theta \neq \theta_0^*$ there exist neighbourhoods $U_{1\psi} \times U_{2\psi}$ of ψ and V_ψ of F_0^* such that

$$
P_{0*} \sup_{\theta \in U_{1\psi}, \theta' \in U_{1\psi}, F \in U_{2\psi}, F_2 \in V} \log \frac{p_\theta(x \mid F)\, p_{\theta'}(y \mid z)}{p_{\theta_0}(x \mid F)\, p_{\theta_0}(y \mid z)} < 0.
$$

By our choice of the maximum likelihood estimator for the retrospective model, (2.1) remains valid, and the proof can proceed as before.

REFERENCES

1. J. M. Begun, W. J. Hall, W. M. Huang, and J. A. Wellner, Information and asymptotic efficiency in parametric-nonparametric models, *Ann. Statist.* **11** (1983), 432–452.
2. R. M. Dudley, A course on empirical processes, *in* "Lecture Notes in Mathematics," Vol. 1097, pp. 1–142, Springer-Verlag, Berlin, 1984.

3. R. D. Gill and A. W. van der Vaart, Non- and semi-parametric maximum likelihood estimators and the von Mises method, Part II, *Scand. J. Statist.* **20** (1993), 271–288.
4. E. Giné and J. Zinn, Lectures on the central limit theorem for empirical processes, *in* "Lecture Notes in Mathematics," Vol. 1221, pp. 50–113, Springer-Verlag, New York/Berlin, 1986.
5. R. Z. Hasminskii and I. A. Ibragimov, On asymptotic efficiency in the presence of an infinite dimensional nuisance parameter, *in* "Lecture Notes in Mathematics" (Ito and Prohorov, Eds.), Vol. 1021, pp. 195–229, Springer-Verlag, New York, 1983.
6. J. Kiefer and J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters, *Ann. Math. Statist.* **27** (1956), 887–906.
7. A. N. Kolmogorov and V. M. Tikhomorov, Epsilon-entropy and epsilon-capacity of sets in function spaces, *Amer. Math. Soc. Transl. Ser. 2* **17** (1961), 277–364.
8. S. A. Murphy and A. W. Van der Vaart, Semiparametric likelihood ratio inference, *Ann. Statist.* **25** (1997), 1471–1509.
9. S. A. Murphy and A. W. Van der Vaart, Observed information in semiparametric models, *Bernoulli* **5** (1999), 381–412.
10. M. Ossiander, A central limit theorem under metric entropy with L_2-bracketing, *Ann. Probab.* **15** (1987), 897–919.
11. D. Pollard, "Convergence of Stochastic Processes," Springer-Verlag, New York, 1984.
12. D. Pollard, "Empirical Processes: Theory and Applications," Institute Mathematical Statistics, Hayward, CA, 1990.
13. J. M. Robins, A. Rotnitzky, and L. P. Zhao, Estimation of regression coefficients when some regressors are not always observed, *J. Amer. Statist. Assoc.* **427** (1994), 846–866.
14. K. Roeder, R. J. Carroll, and B. G. Lindsay, A semiparametric mixture approach to case-control studies with errors in covariables, *J. Amer. Statist. Assoc.* **91** (1996), 722–732.
15. W. Rudin, "Functional Analysis," McGraw–Hill, New York, 1973.
16. A. W. van der Vaart, On differentiable functionals, *Ann. Statist.* **19** (1991), 178–204.
17. A. W. van der Vaart, Maximum likelihood estimation with partially censored observations, *Ann. Statist.* **22** (1994), 1896–1916.
18. A. W. van der Vaart, Efficiency of infinite dimensional *M*-estimators, *Statist. Neerlandica* **49** (1995), 9–30.
19. A. W. van der Vaart, On a model of Hasminskii and Ibragimov, *in* "Proceedings, Probability Theory and Mathematical Statistics, Kolmogorov Semester, Euler International Mathematical Institute, St. Petersburg, 1993" (Zaitsev, Ed.), pp. 297–308, Gordon & Breach, Amsterdam, 1996a.
20. A. W. van der Vaart, New Donsker classes, *Ann. Probab.* **24** (1996b), 2128–2140.
21. A. W. van der Vaart, Efficient estimation in semiparametric models, *Ann. Statist.* **24** (1996c), 862–878.
22. A. W. van der Vaart and J. A. Wellner, "Weak Convergence and Empirical Processes," Springer-Verlag, New York, 1996.